

因果关系论视角下生成式人工智能提供者刑事责任研究

尹 龙

北京师范大学法学院, 北京

收稿日期: 2025年7月18日; 录用日期: 2025年8月8日; 发布日期: 2025年8月20日

摘 要

生成式人工智能快速发展的过程中带来了诸如法益侵害、冲击传统刑事责任主体理论、刑事责任认定困难等刑法困境, 从犯罪论和刑罚论的角度进行分析, 应当否定生成式人工智能的刑事责任主体资格。立足于风险社会和积极主义刑法观的理论背景, 为有效控制和预防生成式人工智能带来的风险, 应当将作为“风险源”的提供者作为归责中心。承担刑事责任需要具备客观和主观两方面的因素, 客观方面的因素关键在于提供者行为和危害结果之间的因果关系判断, 条件说、原因说和相当因果关系有扩大提供者处罚范围和认定标准缺乏统一标准的缺陷, 而客观归责理论分为归因和归责两个阶段, 依据提供者是否创造和实现“法不容许的危险”进行判断, 能够对提供者行为和危害结果之间的因果关系得出妥当结论, 为刑事责任提供稳定的客观依据。

关键词

因果关系, 生成式人工智能, 提供者, 客观归责理论

Criminal Liability of Generative Artificial Intelligence Providers from the Perspective of Causation Theory

Long Yin

School of Law, Beijing Normal University, Beijing

Received: Jul. 18th, 2025; accepted: Aug. 8th, 2025; published: Aug. 20th, 2025

Abstract

The rapid development of generative artificial intelligence has produced a number of criminal-

law challenges, including infringements of legally protected interests, disruption of traditional doctrines concerning subjects of criminal liability, and difficulties in determining criminal responsibility. An analysis from the perspectives of the doctrine of crime and the doctrine of punishment indicates that generative artificial intelligence itself should be denied even a formal status as a subject capable of bearing criminal liability. Against the theoretical background of the risk society and an activist conception of criminal law, effective control and prevention of the risks posed by generative AI call for centering attribution on the provider, understood as the “source of risk.” Criminal liability presupposes both objective and subjective elements. On the objective side, the core issue is how to assess causation between the provider’s conduct and the harmful result. The *conditio-sine-qua-non* test, *causa* theory, and adequate causation approaches either over-extend the scope of punishment applicable to providers or lack uniform criteria. By contrast, the theory of objective imputation proceeds in two stages—*attribution* and *normative imputation*—and evaluates whether the provider created and realized a “legally impermissible risk.” This framework can yield sound conclusions regarding the causal nexus between provider conduct and harmful outcomes, thereby supplying a stable objective basis for assigning liability.

Keywords

Causation, Generative Artificial Intelligence, Provider, Objective Imputation Theory

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 问题的提出

生成式人工智能在 ChatGPT 发布后进入大众视野，在社会多领域掀起了应用和研究浪潮。与传统的人工智能相比，生成式人工智能不直接受制于人的指令，具有一定的自主决策能力，它是一种深度学习模型，通过大量数据识别、训练，深度学习算法并形成一个由参数构成的神经网络，从而模拟人类的创造过程，依据所获信息进行分析，生成新的文本、图片、视频、代码等。基于此特点，生成式人工智能为医学、教育、法律等多领域带来高效、激活创造力、加速决策等好处。然而不得不承认生成式人工智能的应用与发展存在不可避免的负外部性，对个人信息、网络安全、知识产权等法益产生威胁，并构成犯罪。在这些犯罪背后涉及生成式人工智能提供者、使用者以及生成式人工智能本身三方“主体”，应该由哪一方承担相应的刑事责任。当前学界的争论主要集中在生成式人工智能本身是否应当承担刑事责任，即其是否具备刑事责任主体资格，笔者认为生成式人工智能并不具备刑事责任主体资格。刑法一个重要作用是预防犯罪，在风险社会以及人工智能技术不断发展的背景下，应当以积极的刑法观应对此类犯罪，重视涉生成式人工智能犯罪的预防。生成式人工智能的提供者作为犯罪危险的风险源，判断其是否应当承担刑事责任应是讨论的重点，而判断危害结果由哪个具体行为引起的才能确定刑事责任由谁承担，因此本文从因果关系视角论证提供者是否是承担刑事责任的主体，当然需要明确的是不法层面的因果关系解决的是客观要件的问题，仅是承担刑事责任的要素之一，是否承担刑事责任还需要判断主观要件。

2. 规制生成式人工智能的刑法困境

以 ChatGPT 为代表的生成式人工智能在上游的提供以及下游的使用过程中都存在刑事风险，会对个人信息、著作权、网络安全等法益产生威胁，同时也对刑事责任的归属以及共同犯罪理论提出挑战。以

个人信息法益为例，生成式人工智能最根本的是对大量高质量数据的学习，其作用发挥依托大量数据的供给，在提供阶段研究者需要根据数据对生成式人工智能进行训练，在使用阶段生成式人工智能和用户的交互过程中需要对用户的信息进行分析，因此无论是提供阶段还是在使用阶段，生成式人工智能都存在侵犯个人信息的问题。分析生成式人工智能可能带来刑事风险，并从控制风险源角度探讨提供者的刑事责任有利于实现刑法在生成式人工智能犯罪领域的预防作用。

2.1. 生成式人工智能提供阶段产生的法益侵害危险

生成式人工智能在提供阶段主要面临侵害数据安全、个人信息安全等法益的风险。生成式人工智能是依托于对大规模数据的学习，通过算法模型，创造新内容的 AI 技术。因此研究者在对生成式人工智能进行开发、训练都需要海量的数据作为支撑。作为开发、训练的数据来源主要有以数据爬虫技术为典型的主动构建形式和“输入 - 收录”为典型的被动构建形式两种[1]。

其中依赖数据爬虫技术主动构建数据参数过程中，提供者的获取数据的行为违背数据收集的正当性规则，造成对数据法益的侵犯。以 ChatGPT 为例，其作为工具是基于 Transformer 的人工智能应用，主体架构遵从“语料体系 + 预训练 + 微调”的基本模式。其中语料体系包括预训练语料与微调语料两个部分，预训练语料包括 OpenAI 从书籍、杂志、百科、论坛等渠道收集，并初步清理后形成的海量无标注文本数据[2]；微调语料包括从开源代码库爬取、专家标注、用户提交等方式收集和加工的高质量有标注文本数据。通过对海量文本数据的学习以及偏好训练，ChatGPT 才能根据使用者的需求，生成高度符合使用者需求的全新内容。然而在此过程中，OpenAI 并未公布其数据收集来源，“黑箱”问题严重。因此存在生成式人工智能服务提供者将被保护的数据信息内容用于训练人工智能的情形[3]，使数据安全法益受到侵犯，这种侵害数据安全的行爲，可能构成我国刑法第 285 条第二款规定的非法获取计算机信息系统数据罪。依赖“输入 - 收录”这种被动构建数据参数的过程中，生成式人工智能可能涉及违规搜集、存储、泄露用户信息，侵犯个人信息安全法益。生成式人工智能将在信息数据世界的不断扩张和信息数据安全保护之间造成难以弥合的间隙[4]。

2.2. 生成式人工智能使用阶段产生的法益侵害危险

在使用阶段存在使用者滥用生成式人工智能犯罪侵害他人法益以及生成式人工智能“自发”侵害法益的危险。生成式人工智能具有“深度伪造”的功能，能够通过深度学习算法，根据目标人物的音视频资料伪造出高逼真度的视听资料，生成的内容使人信服。[5]例如宁夏彭阳发生了一起利用生成式人工智能实施诈骗的案例，犯罪分子用 AI 换脸和拟声技术，用被害人小龙表哥的头像和声音与其进行交流，使其上当；白某某通过利用 AI “一键去衣”技术，深度伪造淫秽图片近 7000 张，以每张 1.5 元的价格卖出，获利将近一万元，构成传播淫秽物品罪。此外，生成式人工智能还存在“自发”侵犯法益的风险，由于生成式人工智能具有自主学习能力，可能存在超出提供者预期的违法行为，主要体现在其可能存在“自发”地非法获取、分析、操控数据的危险[6]。例如生成式人工智能在更新数据库时，是根据算法获取其可能接触到的一切数据，而无法甄别数据的公开与否，可能存在“自发”非法获取数据的情形。

2.3. 生成式人工智能对刑事责任认定的挑战

根据上述从提供阶段和使用阶段两方面对生成式人工智能带来的法益侵害危险的分析，可以发现涉生成式人工智能犯罪对刑事责任的影响主要在于使责任主体多元化，犯罪背后涉及的主体可能包括提供者、使用者以及生成式人工智能本身。使用者单方利用人工生成式人工智能犯罪的，自然由使用者承担刑事责任，问题的关键在于提供者刑事责任该如何认定。对于结果犯的刑事责任的认定不仅需要存在危害行为与危害结果，而且要求二者之间具有引起和被引起的因果关系。在生成式人工智能犯罪的背景下，

提供者的行为和危害结果之间可能存在介入因素，导致因果关系判断存在困难，为此本文承袭对危险分析的划分，也从生成式人工智能提供和使用两个阶段分析提供者的行为与犯罪结果是否存在因果关系，判断其是否存在承担刑事责任的基础。在提供阶段分为两种情形：(1) 提供者本身存在违法行为，造成法益侵害结果；(2) 提供者不存在违法行为，但仍然发生法益侵害结果。在使用阶段分为两种情形：(1) 生成式人工智能被作为犯罪工具：① 提供者提供生成式人工智能的目的是实施犯罪，投入使用后，使用者利用其犯罪，并最终产生法益侵害实施；② 提供者提供生成式人工智能的目的是促进发展，投入使用后，被使用者恶意利用实施犯罪，产生法益侵害结果；(2) 提供者提供生成式人工智能的目的是促进发展，但生成式人工智能在自我学习更新或运行过程中实施“自发”犯罪。

3. 生成式人工智能刑事责任主体讨论

3.1. 否定生成式人工智能刑事责任主体资格

生成式人工智能带来的刑事风险之一是刑事责任主体多元化，刑事责任的归属产生困惑。特别是针对生成式“自发”实施无需使用者或提供者干预的刑法上侵害法益的行为这一新型犯罪形态，刑事责任该如何认定，为此学界就人工智能能否成为刑事责任主体展开激烈讨论，出现肯定说和否定说之争。

持肯定说的学者认为，可以根据人工智能的“智能”程度，也即是否具有自主的辨认能力和控制能力，将人工智能划分为弱人工智能和强人工智能，否定弱人工智能具有刑事责任主体资格，肯定强人工智能的刑事责任主体资格，肯定的原因主要有：第一，生成式人工智能的出现预示着强人工智能时代“技术奇点”的出现，需要事先明确其刑事责任主体资格，以预防将来可能出现的强人工智能对法益造成的危险。第二，强人工智能具有自主决策能力，具有控制和辨认能力，有必要将人工智能纳入刑事责任主体的范畴进行规制。持否定说的原因有：第一，从人工智能的发展水平看，当前的生成式人工智能仍属于弱人工智能，并没有达到人类所具备的自主性，只是一种自动化程度较高的程序。第二，从刑罚角度来看，对人工智能予以处罚不能达到刑罚造成痛苦的目的。第三，容易形成行为人逃脱刑事责任的途径。

本文对生成式人工智能刑事责任主体资格持否定观点。从犯罪论角度看，从行为理论的变迁与批判过程中，可以发现纯粹的自然行为和纯粹的价值行为都被摒弃，只有综合事实和价值两者的判断才能对行为作出正确判断，因此刑法上的危害行为应当具有有体性和有意性两个特征，从而将单纯的动物行为与观念行为排除在危害行为之外。笔者认为，生成式人工智能的运行过程到产出所包含的“行为”是一种纯粹的自然行为，不具备有意性的特征，不属于刑法危害行为的范畴。其次，生成式人工智能不具备辨认和控制的刑事责任能力以及违法认识可能性。辨认能力是行为人对自己的行为在刑法上的性质和后果的认识能力，而生成式人工智能将自然语言转化为计算机语言，通过算法再生成自然语言，这一过程是纯粹的数据加工过程，它无法通过计算机语言或者自然语言理解背后的含义，不具备感知和理解能力[7]，例如它不能像人类作者一样，确定自己生成的内容无法进行是否包含“淫秽”内容的价值判断。辨认能力是控制能力的基础，没有辨认能力自然也就没有讨论控制能力的必要。从刑罚论的角度来看，对生成式人工智能处以刑法不符合刑罚正当化根据。首先，不符合预防犯罪这一刑罚目的。不能实现一般预防的目的，对某个具体的生成式人工智能处以刑罚，并不能对其他生成式人工智能起到威慑作用；不能实现特殊预防目的，特殊预防是为了使犯罪行为人不复再犯，而生成式人工智能不具有辨认和控制能力，归根结底只能作用工具被人所利用，所以对生成式人工智能进行刑罚处罚，其是否会继续实施犯罪行为最终取决于提供者和使用者，而不取决于它本身。其次，不符合报应正当性这一根据。刑罚种类包括生命刑、自由刑、财产刑三种，而生成式人工智能不具有感知痛苦的能力，因此对其适用生命刑和自由刑没有意义；人工智能也不具有财产，因此也无法对其适用财产型。可能的处罚手段有销毁、暂停使用等，但这种处罚实际上是对其使用者和所有者的处罚，处罚对象最终回到了人这一群体。

3.2. 将生成式人工智能提供者作为归责中心

3.2.1. 符合生成式人工智能时代对刑法有效供给的需求

生成式人工智能的快速发展和应用，带来新型犯罪以及刑事风险，对传统刑法理论特别是对刑事责任主体的传统观点产生强烈冲击，深刻动摇传统刑法体系和基础。面对人工智能时代产生的刑法困境，有学者指出注重事后惩罚的刑法在人工智能时代显得捉襟见肘，需要在人工智能领域建立积极刑法规范，发挥刑法的预防功能[8]。积极主义刑法观使得刑法价值转向预防，并产生预防刑法。预防刑法以风险社会为基础，国家重视通过刑法实现对社会风险的预防与控制。生成式人工智能犯罪背景下，将作为危险源的提供者作为归责中心符合生成式人工智能时代对刑法有效供给的需求。第一，袁彬教授认为生成式人工智能最初的程序是体现提供者意志的，提供者研发的初始程序是生成式人工智能的高智能性的基础[9]。所以作为一种算法模型，生成式人工智能的运行程序以及学习的数据参数来源于提供者，究其本质最终体现的是提供者的选择，因此提供者在生成式人工智能提供阶段的数据学习和使用阶段的功能发挥其根本作用。生成式人工智能投入社会市场中，面对不特定的使用群体，容易引发犯罪，在这种社会风险面前，应当着重讨论提供者的刑事责任，将其作为归责中心，有利于实现对风险的预防和控制。第二，以提供者作为归责中心，有利于发挥一般预防和特殊预防的刑法效果。就一般预防而言，就提供者的刑事责任设立预防刑法，能够唤醒和强化提供者对法的忠诚、对法秩序的贯彻力的信赖，使提供者的法意识安定化[10]。因此，在生成式人工智能犯罪领域，应当区分生成式人工智能犯罪的不同阶段不同情形，妥善解决刑事责任归属问题，将提供者的刑事责任作为讨论重点。

3.2.2. 符合罪责刑相适应原则的要求

《刑法》第五条规定：“刑罚的轻重，应当与犯罪分子所犯罪行和承担的刑事责任相适应。”其中包含两个层面：一是刑罚的轻重应当与犯罪的性质以及严重程度相适应，重罪重罚，轻罪轻罚；二是刑罚的轻重应当与刑事责任的轻重相适应。所以罪刑相适应原则要求对行为人的罪行和刑事责任准确判断。在上文提到提供阶段的提供者本身存在违法行为，导致发生法益侵害结果的情形下，例如搜集数据时不合法不合规，侵害数据安全和个人信息法益的行为，就应查明数据来源并且将法益侵害归属于提供者。使用阶段生成式人工智能被使用者当成工具实施犯罪的情形下，对使用者当然按照实行犯的规定进行归责和处罚，对提供者则应当根据提供者的提供行为或者未尽管理义务的行为与法益侵害结果之间是否存在因果关系，是否与使用者共谋犯罪等因素进行综合判断，合理认定提供者的刑事责任。使用阶段提供者提供生成式人工智能的目的是为了促进发展，但生成式人工智能在自我学习更新或运行过程中实施“自发”犯罪的情形下，例如生成式人工智能在后续的“再学习”时，自动生成的内容侵害了国家、社会或个人法益，该法益侵害后果应该由谁承担。应当认为在这种情况下，生成式人工智能造成的法益侵害的根本在于生成式人工智能内部出现问题或漏洞，最终的责任归属只能是提供者。但责任的分配在考虑法益侵害结果的基础上，还应当综合判断是否存在技术局限等问题，例如现阶段的技术是否能够有效防止生成式人工智能“自发”犯罪。所以在不同犯罪情形下，提供者应当对哪些法益侵害结果承担刑事责任需要具体判断，进而保障对提供者的判处的刑罚的严厉程度不超过提供者应负的刑事责任的程度。综上，将生成式人工智能提供者作为归责中心符合罪责刑相适应原则的要求，有利于实现刑事责任的正确划分，将提供者的刑事责任限定在合理的范围之内。

4. 因果关系视角下生成式人工智能提供者刑事责任之辩证

生成式人工智能提供者在涉生成式人工智能犯罪，侵害法益时是否承担刑事责任，要从主客观两个方面进行判断，对不法层面的判断核心在于判断提供者的行为和危害结果之间是否具有因果关系，如果

具有因果关系则进一步判断提供者主观上是否具有非难可能性，如果不具有因果关系，从客观层面就已经可以排除提供者对危害结果的刑事责任，由此可见，对因果关系的判断是认定提供者刑事责任的第一步。但是学界关于因果关系存在不同的学说，以何种学说为标准认定生成式人工智能犯罪中的因果关系具有合理性值得讨论。

4.1. 因果关系于刑事责任该当性的影响

刑法上的因果关系是指危害行为和危害结果之间的一种引起与被引起的客观存在的关系，张绍谦教授指出刑法因果关系既是客观存在的事实因果关系又是法律要求的法律因果关系，它们是作为刑事责任的客观根据而存在于刑法之中的[11]。我国刑法秉持罪责自负原则，罪责自负“意味着国家在进行刑事责任归属，进而作出刑罚处罚时，不能将他人应负的责任归咎于特定的个人，同时，也不能将犯罪人应承担的刑事责任转嫁给第三人。”[12]罪责自负原则表明行为人只对自己的危害行为造成的危害结果承担刑事责任，因此判断生成式人工智能提供者行为与法益侵害结果之间是否存在因果关系，是决定提供者对法益侵害结果是否负刑事责任的客观依据。作为判断刑事责任的客观依据，因果关系会影响刑事责任的有无和刑事责任的程度，判断因果关系对解决刑事责任问题意义重大。

因果关系影响刑事责任的有无。刑法上因果关系解决的是危害结果的归属问题，如果行为与结果之间存在因果关系，就意味着该结果能够归属于行为人，即只有具有因果关系才能具备使行为人承担刑事责任的客观责任基础。第二，因果关系影响刑事责任的程度。在一个犯罪过程中，有的危害行为对结果的产生起着直接和决定性的作用，有的危害行为对结果的产生起间接和非决定性作用，因此不同程度的危害行为对应的客观责任也应当有程度上的差别。

4.2. 各学说下生成式人工智能提供者刑事责任的辩证分析

上文划分的四种情形，即情形一：提供阶段，提供者本身存在违法行为，造成法益侵害结果；情形二：提供阶段，提供者不存在违法行为，但仍然发生法益侵害结果；情形三：使用阶段，生成式人工智能被作为犯罪工具：① 提供者提供生成式人工智能的目的是为了实施犯罪，投入使用后，使用者利用其犯罪，并最终产生法益侵害结果；② 提供者提供生成式人工智能的目的是促进发展，投入使用后，被使用者恶意利用实施犯罪，产生法益侵害结果；情形四：使用阶段，提供者提供生成式人工智能的目的是为了促进发展，但生成式人工智能在自我学习更新或运行过程中实施“自发”犯罪。在以上四种情形中情形一和情形三中的第一种情况中提供者存在明显的危害行为且与危害结果的产生具有直接的因果关系，而情形二、情形三的第二种情况以及情形四中提供者的行为和最终产生的法益侵害结果之间的因果关系难以认定，不同的学说也可能产生不同的结论，下文将重点对这几种情形展开讨论。

4.2.1. 条件说扩大对生成式人工智能提供者处罚范围

条件说立足于逻辑的因果关系立场，认为一切行为只要是在逻辑上是引发结果发生的条件，就是结果发生的原因，也即存在“如果没有前者，就没有后者”的关系时，就存在刑法上的因果关系，因此也被称为“全条件同价值说”[13]。条件说的可见条件说并不区分引发结果发生的各个条件的作用和价值大小，甚至可以理解为各个条件都是等价值的。条件说在因果关系的判断上采取的是“假定消除公式”这一方法论，即“非 P 则非 Q”，认为在如果没有前行为，后结果也会发生的情形下，因果关系就被否定；在如果没有前行为，后结果就不会发生的情形下，就可以肯定因果关系存在。在这种方法论下，因果关系的判断就成为完全的事实判断[14]。因为条件说将所有物理上的条件都作为刑法上的因果关系，因而遭到了学界的批判，认为条件说会扩大因果关系的范围，使得刑事责任的客观基础变得过宽[15]。在生成式人工智能犯罪的领域，按照条件说“非 P 则非 Q”的判断标准，认为如果没有提供者提供生成式人工智能，

就不会产生后续的生成式人工智能参与造成的法益侵害结果，因此以条件说判断上述四种情形中提供者的刑事责任，可以得出在所有情形下提供者的行为和法益侵害结果之间均具有因果关系的结论，显然这种结论不具有合理性。例如情形三的第二种情况，在提供者客观上不存在危害行为，主观上不具有罪过的前提下，使用者滥用生成式人工智能实施犯罪活动，仅仅因为提供者的提供行为是使用者使用的前提就将使用者侵害法益的结果归属于提供者，就不当扩大了对提供者的处罚范围，使提供者为他人的行为承担刑事责任。这一方面违背了罪责自负原则，另一方面只要生成式人工智能产生侵害法益的情况，就将原因溯及到提供者，不利于实现科技发展和刑法规制之间的平衡，制约生成式人工智能领域的创新发展。因此不能依据条件说判断提供者行为和法益侵害结果之间的因果关系。

4.2.2. 原因说标准不一导致提供者因果关系认定随意性

原因说是在进一步限制条件说的基础上提出的，原因说在条件说的基础上，主张在导致结果发生的条件中根据一定的标准选出导致结果发生的原因，将其他条件排除在原因之外。但是学者对这一学说也进行了批判，张明楷教授认为从造成结果的众多条件中挑选一个条件作为原因是困难和不现实的，会导致因果关系认定的随意性^[10]。梅向华教授认为原结果的发生并非总是依赖于一个单纯的条件，应当承认复数条件竞合为共同原因，原因说最终没有摆脱条件说的缺陷^[16]。刘艳红教授指出从众多条件中选取原因又存在优势条件说、最有力条件说、动力条件说、最终条件说等不同学说。认为原因说的根本缺陷在于，对于结果发生的诸多条件的筛选标准具有一定的主观性和不确定性，充满主观性的价值判断，得出的“原因”条件的结论不稳定^[17]。因此对原因的选取如果采取不同的学说就会得出不同的结论。例如对于上述情形二，按照优势条件说，对结果发生起决定作用的条件是原因，则认为提供阶段即使提供者不存在违法行为，提供者在提供阶段的行为和侵害数据安全、个人信息法益的侵害结果就不具有因果关系；按照重要条件说，对结果发生具有重要作用的条件就是原因，则认为提供者设计爬虫程序的行为对侵害数据安全、个人信息法益的危害结果具有重要作用，应当认为这种情况下提供者的行为和危害结果之间具有因果关系。对于上述情形三中的第二种情况，即提供者提供生成式人工智能的目的是为了促进发展，投入使用后，被使用者恶意利用实施犯罪，产生法益侵害结果。若按照最终条件说，时间上和结果发生最近的条件是原因，则使用者恶意利用生成式人工智能实施犯罪的行为是和法益侵害结果在时间上距离最近，因此使用者的恶意利用行为是原因，而提供者的提供行为只是条件从而排除在刑法上原因之外；按照最有力条件说，对于结果的发生最为有力的条件是原因，则认为生成式人工智能的技术优势对结果的发生是最有力的条件，因此提供者提供生成式人工智能的行为和法益侵害结果之间具有因果关系。因此，若采取原因说，对原因选取的根据没有统一的标准，对同一种情形，对生成式人工智能提供者行为和侵害结果之间因果关系的判断会得出不同的结论，从而对其是否承担、在什么范围内承担刑事责任得出不同的结论，造成认定的随意性，导致司法认定的混乱。

4.2.3. 相当因果关系说缺乏相当性判断的具体标准

相当因果关系说的核心在于“相当性”，即根据社会生活的一般经验，从行为人行为时来看，某种行为产生某种结果被认为是相当的，就能肯定行为和结果之间具有因果关系。反之，如果行为导致该结果发生是异常的、罕见的则可以否定行为和结果之间的因果关系。对于相当性的判断基础，理论上存在主观说、客观说和折中说三种观点。主观说以行为人的主观认识，即行为人已经认识到或可能认识到的事实作为相当性判断的基础，客观说重视对介入因素是否异常进行判断，以行为时的一切客观事实作为判断基础，折中说则以行为时一般人的认识和行为人的特殊认识到的事实作为判断基础，在三种学说中客观说为主流观点。

根据相当因果关系说中的客观说，在上述生成式人工智能犯罪的四种情形中，认定提供者行为和法

益侵害结果之间的因果关系，关键在于对有无介入因素以及介入因素是否异常的判断，从而判断是否属于介入因素中断因果关系、多因一果的情形。(1) 在情形一中，提供阶段提供者本身存在违法行为造成法益侵害结果，例如提供者利用爬虫程序，违背数据收集的正当性规则获取数据，侵害数据安全，这一情形下，提供者的行为和法益侵害结果之间不存在介入因素，提供者的行为是引发法益侵害结果的原因。(2) 在情形二中，提供阶段提供者不存在违法行为，但仍然发生法益侵害结果，例如生成式人工智能在通过“输入-收录”这一形式获取用户信息侵害个人信息法益时，提供者虽然未违背数据获取规则，没有为生成式人工智能植入不当程序，但是也没有设置响应程序防止生成式人工智能不当获取个人信息，因此生成式人工智能不当获取这一“行为”不是异常介入因素，不能阻断提供者的行为和法益侵害结果之间的因果关系。(3) 在情形三的第一种情况中，提供者提供生成式人工智能的目的是为了实施犯罪，提供者在生成式人工智能中植入违法程序，投入使用后使用者利用其犯罪并最终产生法益侵害结果，使用者的利用行为和提供者的违法行为共同引发法益侵害结果，应当认定二者的行为与结果的发生均具有因果关系，属于“多因一果”。(4) 在情形三的第二种情况中，提供者提供生成式人工智能的目的是为了促进发展，提供者在生成式人工智能中没有植入违法程序，投入使用后被使用者恶意利用实施犯罪，产生法益侵害结果，此时使用者的恶意利用行为属于异常的介入因素，且对危害结果的发生起决定作用，阻断了提供者的行为与法益侵害结果之间的因果关系。(5) 在情形四中提供者提供生成式人工智能的目的是为了促进发展，但生成式人工智能在自我学习更新或运行过程中实施“自发”犯罪的行为是否属于异常的介入因素，存在不同的观点。笔者认为，需要结合提供者是否履行监管义务进行判断，如果提供者在生成式人工智能自我更新时履行监管义务但由于技术原因未能防止的，这种“自发”犯罪显然违背提供者的初衷，就属于异常的介入因素，阻断提供者的行为和法益侵害结果之间的因果关系；如果提供者在生成式人工智能自我更新时没有履行监管义务，这种“自发”犯罪就不属于异常的介入因素，提供者的行为和法益侵害结果之间具有因果关系。即在情形一、情形二、情形三的第一种情况下，提供者的行为和法益侵害结果之间具有因果关系；情形三的第二种情况下，提供者的行为和法益侵害结果之间不具有因果关系；情形四中需要分情况判断因果关系。但相当因果关系中的客观说存在弊端，即判断介入因素是否异常缺乏适当具体的标准，“相当性”的基准不明确。因此无法根据此学说判断生成式人工智能犯罪领域提供者的行为和结果之间的因果关系。

4.3. 客观归责理论妥当性之证成

客观归责理论是由德国罗克辛教授提出的，是指“一个由行为人所因果性地引起的结果，只有当行为人的行为针对行为对象，创设了一个不被容许的危险，并且这些危险在具体的结果中实现时，才能将这个结果归属于客观构成要件。”^[18]基于此，判断一个行为进行归责时需要具备三个条件：第一，制造了法不允许的风险；第二，该实现了法所不允许的风险，风险现实化为结果；第三，该结果在构成要件的保护范围之内。

张明楷教授认为客观归责理论是以事实的因果关系为前提得规范评价理论，也是构成要件理论。我国应当在构成要件部分借鉴客观归责理论的具体内容，在“因果关系”部分应当借鉴危险实现的基本内容，讲因果关系理解为广义上的因果关系即包括事实的因果关系和结果归属两部分，分别进行事实和价值判断^[18]。本文基于因果关系角度探讨生成式人工智能提供者刑事责任客观依据，即其不法性，因此本文将客观归责理论中的“归责”限定在客观阶层，属于不法的范畴，不掺杂提供者主观的判断。

根据客观归责理论对归因和归责的区分，对生成式人工智能提供者的行为和法益侵害结果之间的因果关系也要从归因和归责两个层面分析，首先按照条件说，没有提供者的提供行为，就不会产生后续生成式人工智能侵害法益的结果，因此提供者的行为和危害结果之间具有事实上的因果关系；下一步则需

要从规范的角度判断能否进行归责，即根据上述三个条件“制造风险－实现风险－风险保护”进行具体判断。下面将对上文提到的四种情形具体分析。

情形一：提供阶段，提供者本身存在违法行为造成法益侵害结果。生成式人工智能即使具有自我更新和学习能力，但其本质上依然是依靠算法进行运作的，提供者需要为其输入代码、收集数据，预先设定生成式人工智能的算法。在此阶段提供者本身在收集数据过程中存在违法行为，增加了数据安全被侵害的风险，创设了法不允许的风险。生成式人工智能在学习提供者收集到的数据时，就已经现实侵害了数据安全甚至是个人信息安全，形成法益侵害结果，风险现实化为结果。同时该侵害数据安全、侵害个人信息的结果没有超出构成要件的保护范围。因此，在情形一中，提供者的行为和结果之间具有因果关系。

情形二：提供阶段，提供者表面不存在违法行为，但仍然发生法益侵害结果。根据《生成式人工智能服务管理暂行办法》第四条规定：“提供和使用生成式人工智能服务，应当遵守法律、行政法规，尊重社会公德和伦理道德，遵守以下规定：……(四)尊重他人合法权益，不得危害他人身心健康，不得侵害他人肖像权、名誉权、荣誉权、隐私权和个人信息权益；……”，可见提供者未设置防止生成式人工智能侵犯个人权益的程序就属于法不容许的危险。提供者虽然不存在违背数据收集归责的作为的违法行为，但是提供者作为生成式人工智能的管理方没有设置防止生成式人工智能在与使用者交互的过程中违法获取使用者信息的相应程序，倘若提供者设置了相应程序，就能避免不当收集使用者信息的，但因为其没有设置而未能避免，就是创设了法不容许的风险，并最终侵害使用者的个人信息安全法益，这种情况下提供者的不作为和法益侵害结果之间存在关联，因此在情形二中，提供者的不作为和结果之间具有因果关系。

情形三的第一种情况：提供者提供生成式人工智能的目的是为了实施犯罪，投入使用后使用者利用其犯罪，并最终产生法益侵害结果。提供者就是为了实施犯罪而在生成式人工智能中植入犯罪程序的，比如在系统中植入专门生成淫秽物品的程序，在植入犯罪程序时就已然创设了法不允许的危险，投入使用后，使用者利用其犯罪就将危险现实化为危害结果，应当肯定提供者恶意植入犯罪程序的行为和危害结果之间具有因果关系。

情形三的第二种情况：提供者提供生成式人工智能的目的是为了促进发展，投入使用后被使用者恶意利用实施犯罪，产生法益侵害结果。此情况中，提供者的目的是为了促进发展，并没有创设法不允许的危险，应当否认提供者的行为和危害结果之间的因果关系。或者说提供者的行为本身就不是“危害行为”，从根本上达不到因果关系中“因”的要求。

情形四：提供者提供生成式人工智能的目的是促进发展，但生成式人工智能在自我学习更新或运行过程中实施“自发”犯罪。在这种情况下，提供者一般都是生成式人工智能领域的专家，对风险就有较强的管理能力，应当承担起监管义务，对生成式人工智能保留控制权。如果提供者没有履行监管义务，也就是放弃了对生成式人工智能的控制权，使生成式人工智能“自发”具备可能性，创设了法不允许的危险，并且在生成式人工智能自我学习更新过程中使危险转化为现实，此时提供者未履行监管义务的行为和危害结果之间就具有因果关系；反之则不具有因果关系。

根据客观归责理论，对提供者的行为和危害结果之间的因果关系判断分两步走：第一步，先在条件说的基础上进行“归因”，认为生成式人工智能提供者的行为和危害结果之间具有事实上的因果关系；第二步，根据“制造风险－实现风险－风险保护”三个条件进行“客观归责”判断，从规范和价值判断上对“归因”结果进行检验。通过两个阶段的判断可以对提供者的行为和危害结果之间的因果关系进行准确合理的认定，从而为判断提供者的刑事责任提供稳定的客观依据。

5. 结语

本文基于否定生成式人工智能刑事责任主体资格的前提下，将提供者作为归责中心。但是在生成式

人工智能领域，需要考虑技术发展和刑法规制的平衡问题。如果不予以刑法规制，将会面临新型犯罪问题突出，社会安定性遭到巨大破坏的风险；如果刑法规制超过一定限度会使提供者畏首畏尾，不敢创新，制约该领域的技术创新和发展。因此需要合理界定提供者行为和危害结果之间的因果关系，进而准确认定提供者的刑事责任。刑法中的因果关系学说主要有条件说、原因说、相当因果关系说、客观归责理论等，其中条件说会扩大对提供者的处罚范围，例如在提供者主客观上均合理正当的前提下，将使用者滥用生成式人工智能实施犯罪活动的危害结果归属于提供者，就使提供者为用户的行为承担刑事责任，这显然不合理。另外原因说和相当因果关系说都存在标准不一的情况，按照不同的标准会得出不同的结论，无法为刑事责任提供稳定的客观依据。而客观归责理论可以对不同阶段不同情形下提供者的行为和结果之间的因果关系得出妥当的判断，为刑事责任提供稳定的客观依据，进而有利于准确认定提供者的刑事责任，尽可能最大化追求科技发展和刑法规制的平衡。

参考文献

- [1] 中国法院网. 欧盟开发、运用生成式人工智能系统中的数据保护体系概[EB/OL]. 2024-9-60.
<https://www.chinacourt.cn/article/detail/2024/09/id/8102616.shtml>, 2025-03-02.
- [2] 钱力, 刘熠, 张智雄, 李雪思, 谢靖, 许钦亚, 黎洋, 管铮懿, 李西雨, 文森. ChatGPT 的技术基础分析[J]. 数据分析与知识发现, 2023, 7(3): 6-15.
- [3] 环球网. 起诉 OpenAI 和微软,《纽约时报》打响维权第一枪[EB/OL]. 2023-12-29.
<https://world.huanqiu.com/>, 2025-03-02.
- [4] 盛浩. 生成式人工智能的犯罪风险及刑法规制[J]. 西南政法大学学报, 2023, 25(4): 122-136.
- [5] 刘水灵, 董文凯. Sora 等生成式人工智能对诈骗犯罪司法认定的影响与应对[J]. 青少年犯罪问题, 2024(3): 114-124.
- [6] 刘宪权. 生成式人工智能对数据法益刑法保护体系的影响[J]. 中国刑事杂志, 2023(4): 20-34.
- [7] 李翔, 旷银. ChatGPT 类人工智能及其生成物的刑法思考[J]. 贵州师范大学学报(社会科学版), 2023(4): 78-91.
- [8] 李文吉. 论人工智能时代的预防性刑法立法[J]. 大连理工大学学报(社会科学版), 2020, 41(5): 96-103.
- [9] 袁彬, 薛力铭. 生成式人工智能的刑法规制问题研究[J]. 河北法学, 2024, 42(2): 140-159.
- [10] 张明楷. 刑法学[M]. 第6版. 北京: 法律出版社, 2021: 225-679.
- [11] 张绍谦. 刑法因果关系研究[M]. 北京: 中国检察出版社, 2004: 111.
- [12] 郑延谱. 罪责自负原则——历史演进、理论根基与刑法贯彻[J]. 北京师范大学学报(社会科学版), 2014(4): 99-106.
- [13] 陈兴良. 本体刑法学[M]. 北京: 中国人民大学出版社, 2017: 98-260.
- [14] 黎宏. 刑法因果关系论考察[J]. 清华法学, 2022, 16(3): 112-130.
- [15] 钱叶六. 刑法因果关系理论的重要发展与立场选择[J]. 中国刑事法杂志, 2022(4): 95-111.
- [16] 梅象华. 刑法上相当因果关系之经验解读[J]. 河北法学, 2011, 29(3): 130-138.
- [17] 刘艳红. 刑法因果关系理论的横断切面与危险的现实说之确立[J]. 政治与法律, 2024(8): 64-82.
- [18] 刘艳红. 客观归责理论: 质疑与反思[J]. 中外法学, 2011, 23(6): 1216-1236.